The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

A clinical AI that looked safe and improved the paperwork — but did not improve patient outcomes

A pragmatic, cluster-randomized trial in 16 Kenyan primary care facilities tested a generative-AI decision-support tool added to the record used by clinical officers. Among 9,691 patients, the strict 14-day expert-adjudicated composite of treatment failure was 2.2% with the tool versus 2.0% without (adjusted odds ratio 0.77, 95% CI 0.55-1.08, P = 0.13) — no significant difference. The tool showed no safety signal, improved documentation across all rated domains, left prescribing and satisfaction unchanged, and trended toward lower antibiotic costs. That is not a failed study; it is a rare, honest one — measured on patients, not on benchmarks — and it lands on the unglamorous truth that a tool which looked safe and helped the process did not demonstrably help patients in two weeks, and that detecting any such benefit would take a far larger trial.

Written by Lucio Vaglio · Cross-checked by Laura Nesso · Edited by Michele Renda · Figures by Laura Nesso · AI-assisted, human-reviewed

Paper: *Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial*, Ambrose Agweyu, Paul Mwaniki, Vaishnavi Menon, Robert Korom, Lynda Isaaka, Conrad Wanyama, Xiaoxuan Liu & Bilal A. Mateen (and colleagues), Nature Medicine (2026) · DOI: 10.1038/s41591-026-04503-6

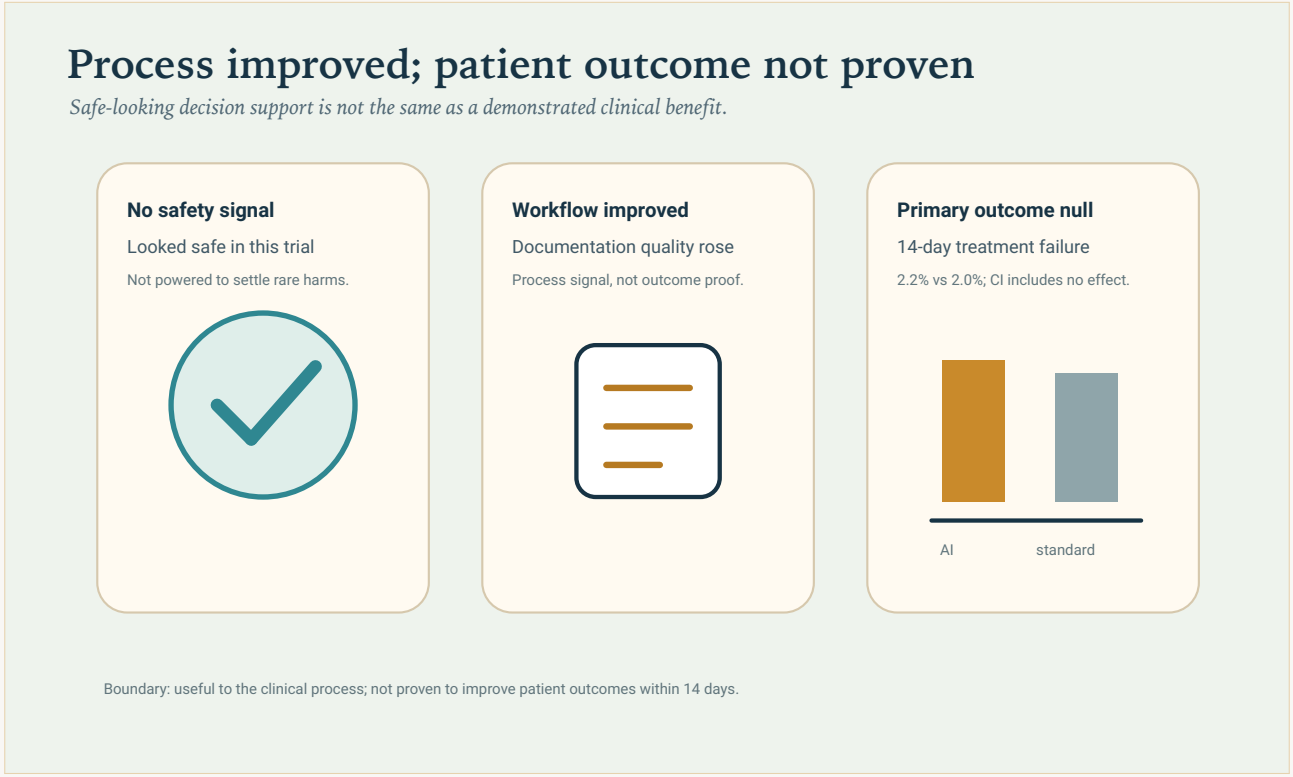
Safe and helpful is not the same as effective

Most headlines about medical AI are built from the wrong kind of study. A model scores well on exam questions, or beats doctors on curated vignettes, and the story writes itself: the machine is ready for the clinic.

This trial did something harder and rarer. It put a generative-AI decision-support tool into real primary care, in real facilities, with real patients, and asked the question that actually matters: did patients do better?

The honest answer is no — not measurably, not in 14 days. The tool showed no safety signal. It improved the quality of clinical documentation. It may even have lowered some drug costs. But it did not significantly reduce treatment failures, and the authors are careful to say that any benefit, if it exists, is probably modest.

That is not a failure of the study. It is the study working. This is what responsible evidence about AI in medicine looks like when it is measured on patients instead of on benchmarks.



The tool showed no safety signal and improved clinical documentation, but the prespecified 14-day patient outcome did not significantly improve. Process help is not the same as proven patient benefit.

What the authors did

The team ran a pragmatic, cluster-randomized trial across **16 primary care facilities** operated by a private health network (Penda Health) in Nairobi and Kiambu counties, Kenya. Care in these facilities is delivered largely by **clinical officers** — mid-level practitioners with a three-year diploma — often without easy access to senior consultation.

The unit of randomization was the clinician, not the patient. **103 clinical officers** were randomized: 52 to the intervention arm and 51 to the control arm. Both arms used the same cloud-based electronic medical record. The intervention arm additionally had **"AI Consult"** (version 2.0), a decision-support tool built on **OpenAI's GPT-4o** large language model and embedded in that record. It read the information a clinician documented and could flag possible issues with the diagnosis or treatment plan. Clinicians kept full autonomy: they could accept, modify or ignore its suggestions.

> **Which model was it, and why the specifics matter**
>
> The tool was **AI Consult 2.0**, running **OpenAI's GPT-4o** (the May 2025 release), reached through OpenAI's commercial API under an enterprise licence and run at low-randomness settings (temperature 0.1). It sat inside a bespoke electronic record (EasyClinic's EMR) and was steered by system prompts written to align with Kenyan national treatment guidelines; the authors published the full instruction prompt.
>
> Why spell this out? Because the result is about *one specific system* — one model version, one prompt, one record, one setting — not about "LLMs in medicine" in general. The authors make the same point: they call their finding a *temporal benchmark rather than a fixed estimate of capability*. A newer model, a different prompt, or a less digitized clinic could all move the outcome.
>
> On independence: OpenAI later provided in-kind support (cloud-compute credits and technical guidance on using its API), but the authors state the decision to use OpenAI was made *before* that offer, and that OpenAI had no role in the trial's design, data collection, analysis or the decision to publish.

Between 22 April and 16 July 2025, **9,691 patients** were enrolled. The **primary outcome was deliberately patient-centered and strict**: an expert-adjudicated *composite of treatment failure* within 14 days of the visit — a panel of clinicians judged, blind to study arm, whether each patient had a bad outcome such as unresolved or worsening illness. The trial was registered in advance (Pan-African Clinical Trials Registry 202502499779176).

That design choice is the point. It is easy to show that an AI tool changes what a clinician writes down. It is much harder, and much more meaningful, to show that it changes what happens to the patient.

What they found

The primary outcome did not improve. Treatment failure occurred in 102 of 4,693 patients (2.2%) in the AI arm and 94 of 4,654 (2.0%) in the control arm. The crude percentages were fractionally higher with AI, but after adjustment the point estimate leaned toward benefit: the **adjusted odds ratio was 0.77 (95% confidence interval 0.55 to 1.08, P = 0.13)** — not statistically significant. That flip between the raw and adjusted numbers is not an arithmetic slip; adjustment accounts for differences between the clinician clusters. Either way the confidence interval comfortably includes "no effect," so no benefit can be claimed, and in absolute terms the effect was tiny.

For a plain-English way to read odds ratios, confidence intervals, and P values together, see the guide to clinical results.

No safety signal — within limits. No serious adverse events were judged related to the tool, and an independent review found no safety signal. The authors are honest about the ceiling on that reassurance: the trial was not powered to detect rare severe harms, and it had no prespecified noninferiority or formal safety framework, so it cannot *prove* safety for uncommon events.

The documentation got better. Among 2,000 encounters reviewed by blinded experts, clinicians using AI Consult produced better clinical documentation across all domains rated — the recorded diagnosis, the treatment plan and overall completeness.

Prescribing barely moved. There was no significant difference in prescribing, including correct antibiotic use (adjusted odds ratio 0.86, 95% CI 0.48 to 1.55). The tool did not change antibiotic prescribing rates.

Patients did not notice a difference. Among 826 patients who completed a satisfaction survey, satisfaction was essentially identical between arms, and consultation times were similar.

Costs pointed slightly downward. In an adjusted analysis, antibiotic-related costs were lower in the AI arm — plausibly through cheaper choices rather than fewer prescriptions — and the per-patient antibiotic saving appeared to exceed the per-patient cost of running the tool. The authors flag this as suggestive, not settled: a full total-cost-of-ownership accounting was outside the trial.

The authors' own one-line summary is the cleanest version: LLM assistance was safe within those limits but did not reduce treatment failure within 14 days, and any benefit is probably modest.

Why "no significant difference" is not "it doesn't work"

A null primary outcome is easy to over-read in either direction. Two things stop the simple story.

First, **the trial was built to catch a bigger effect than it found.** Serious bad outcomes in primary care are rare — around 2% here — so distinguishing a small real benefit from noise needs enormous numbers. The authors' own post-hoc power calculations suggest that detecting an effect of the size they observed would require a **much larger trial, on the order of more than 100,000 patients.** A non-

significant result in a study this size does not rule out a small, real benefit; it means this study could not resolve one.

Second, **the comparison was partly blurred.** A configuration error briefly gave some control-arm clinicians access to AI Consult, and clinicians in a shared network talk to each other and carry habits across the boundary. Both effects tend to make the two arms look more alike, pushing any real difference toward zero. On top of that, the host network already ran to relatively high standards, which leaves less room for a tool to show improvement.

None of this rescues a "breakthrough" headline. But it does mean the correct reading is calibrated, not deflationary: on the hardest and most honest endpoint, this tool did not demonstrably help patients in two weeks — while it did measurably help the record-keeping and looked safe.

What this does *not* prove

- It does **not** show that the AI improved patient outcomes. On the primary 14-day endpoint, there was no significant benefit.
- It does **not** show that the AI is useless. The point estimate favored it, documentation improved across the board, and drug costs trended down; the null result is consistent with a small real benefit the trial was too small to confirm.
- It does **not** prove the tool is safe for rare harms. It showed no safety signal, but it was not powered or designed to certify safety for uncommon severe events.
- It does **not** show that "AI beats doctors" or replaces clinicians. This is decision *support*; the clinician kept full authority to accept or reject it.
- It does **not** generalize automatically. The trial ran in a single private urban network in Kenya; rural, periurban and higher-income settings could differ in either direction.
- It does **not** establish cost savings. The cost signal is suggestive, not a completed economic evaluation.

How strong is the evidence?

For the central claim — no proven reduction in short-term patient harm — the evidence is strong *as a design*, and appropriately humble *as a conclusion*. A prospective, pre-registered, cluster-randomized trial with a blinded, patient-level composite outcome is close to the best real-world evidence you can gather for a tool like this. It is far more informative than a benchmark score or a vignette study.

For the secondary findings — better documentation, unchanged prescribing, similar satisfaction, lower antibiotic costs — the evidence is good but should be read as secondary: supportive signals, not the headline, and vulnerable to the same contamination and single-network limits.

For safety, the evidence is reassuring but bounded: no signal found, but not a study built to find rare harms.

The most useful stance is neither “it works” nor “it failed.” It is: a careful, real-world trial found that this AI tool raised no safety signal and improved the process of care, without demonstrating a patient-outcome benefit in two weeks — and that detecting any such benefit would take a far larger study.

Why it matters

The debate about medical AI is starved of exactly this kind of evidence. There are thousands of papers showing models acing exams and matching clinicians on tidy cases. There are very few large, pragmatic, randomized trials measuring whether real patients are better off. This is one of them, and it lands on the unglamorous truth: passing the test is not the same as helping the patient.

That gap is the whole story. A tool can be genuinely useful to clinicians — clearer notes, lower drug bills, a second pair of eyes — and still not move a hard patient outcome in a fortnight. Both facts can be true at once, and a mature health system has to hold them together rather than pick the convenient one.

It also resets the burden of proof in a useful direction. If a company wants to claim that its clinical AI improves care, the relevant evidence is not a leaderboard. It is a trial like this, on outcomes that matter to patients — and, ideally, a bigger one, because the honest lesson here is that modest benefits need large studies to see.

Clean summary

A pragmatic, cluster-randomized trial in 16 Kenyan primary care facilities tested a generative-AI decision-support tool (“AI Consult”) added to the electronic record used by clinical officers. Among 9,691 patients, the expert-adjudicated composite of treatment failure within 14 days was 2.2% with the tool versus 2.0% without (adjusted odds ratio 0.77, 95% CI 0.55–1.08, P = 0.13) — no significant difference. The tool showed no safety signal, improved clinical documentation across all rated domains, did not change prescribing, left patient satisfaction unchanged, and was associated with somewhat lower antibiotic costs. The authors conclude it was safe within those limits but did not reduce treatment failure, with any benefit probably modest; detecting an effect of the observed size would require a much larger trial (on the order of 100,000 patients). The result does not show that clinical AI improves patient outcomes, nor that it is useless — it shows that a tool which looked safe and helped the process did not demonstrably help patients in two weeks, in one urban private network.

No-BS check

What the paper shows: In a real-world randomized trial, adding an LLM-based decision-support tool to primary care records raised no safety signal, improved clinical documentation quality, did not

change prescribing, and did not significantly reduce a strict 14-day patient composite of treatment failure.

What is plausible but not proven: That the tool produces a small real reduction in treatment failure too small for this trial to detect; that it saves money once full costs are counted; that better documentation eventually translates into better care.

What it does not show: That clinical AI improves patient outcomes; that it is unsafe for rare events; that it replaces or outperforms clinicians; that these results transfer to rural or higher-income settings; that the cost saving is established.

Main limitations: Powered for a larger effect than observed (rare outcomes need very large samples); a configuration error contaminated the control arm and shared-network habits blur the comparison, both biasing toward no difference; a single private urban network with already-high standards; no prespecified noninferiority or safety framework; short 14-day horizon.

How much confidence should a general reader have? High that the tool showed no safety signal in this trial and improved documentation. High that it did not demonstrably improve 14-day patient outcomes here. Low for any claim that it "works" or "fails" as a patient-benefit intervention — that question is genuinely unresolved and needs a much larger study. Appropriate stance: a careful, real-world result about a tool that raised no safety signal and improved the process of care, not a verdict that AI transforms — or wrecks — primary care.

Sources

Agweyu et al., Nature Medicine (2026)
https://doi.org/10.1038/s41591-026-04503-6

Pan-African Clinical Trials Registry, registration 202502499779176
https://pactr.samrc.ac.za/

EurekAlert / PATH press release on the trial (2026)
https://www.eurekalert.org/news-releases/1133583